DOCUMENT RESUME

ED 348 393 TM 018 836

AUTHOR Goldberg, Gail Lynn; Kapinus, Barbara
TITLE Problematic Responses to Reading Performance Assessment Tasks: Sources and Implications.
PUB DATE 3 Apr 92
NOTE 32p.
PUB TYPE Reports - Evaluative/Feasibility (142) -- Tests/Evaluation Instruments (160)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Comparative Testing; Elementary Education; Elementary School Teachers; *Evaluators; Questionnaires; Reading Tests; *Response Style (Tests); *Scoring; State Programs; Student Evaluation; *Test Construction; Testing Problems; Testing Programs
IDENTIFIERS *Maryland School Performance Assessment Program; Open Ended Questions; *Performance Based Evaluation

ABSTRACT

The Maryland School Performance Assessment Program (MSPAP) is a relatively new, statewide performance assessment of students in grades 3, 5, and 8. When first administered in May of 1991, the MSPAP included a battery of performance assessment tasks designed to generate written or drawn responses to reading texts. This study evaluated selected reading activities, responses, and scoring apparatuses to improve performance task development and scoring strategies for future MSPAP editions and for other assessments using constructed responses. Focus was on identifying discrepancies between the range of responses anticipated when activities were developed and the actual response behaviors triggered by some activities, problems inherent in the activities themselves, and problems resulting from a mismatch between responses and the scoring apparatus that had been developed to characterize the range of performance on particular activities. Two major test forms in reading/writing/language usage administered in grades 3, 5, and 8 were studied. Each test form contained approximately 26 brief open-ended questions and 1 extended question that yielded a more developed response to be scored for reading, writing, and language usage. Questionnaire data on these 2 test forms were collected from all 12 teams of trained Maryland teachers serving as readers (N=123). Across the test forms examined, contextual errors (those triggered by the discrepancies between classroom and testing contexts) were rare. A seven-item list of references is included. The readers' questionnaire and scoring rubrics are appended. (RLC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Problematic Responses to Reading Performance Assessment Tasks: Sources and Implications

ED348393

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Gail Lynn Goldberg and Barbara Kapinus

Maryland State Dept. of Education

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
GAIL LYNN GOLDBERG
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

The Maryland School Performance Assessment Program (MSPAP) is a relatively new, statewide performance assessment of students at grades 3, 5 and 8. As first administered in May, 1991, MSPAP included a battery of performance assessment tasks designed to generate written or drawn responses to reading texts. The purpose of this study was to evaluate selected reading activities, responses and scoring apparatus, in order to improve performance task development and scoring strategies: in particular, for future editions of the Maryland School Performance Assessment Program, and in general, for other assessments using constructed response. Our evaluation centered upon discrepancies between the range of responses anticipated when activities were developed and the actual response behaviors triggered by some activities. We sought to identify both those problems inherent in the activities themselves and those resulting from a mismatch between responses and the scoring apparatus that had been developed to characterize the range of performance on particular activities--that is, the sources of cognitive dissonance during scoring.

TM018896

BEST COPY AVAILABLE

Task development for the reading portion of the 1991 Maryland School Performance Assessment Program was conducted by Maryland specialists, administrators and teachers with expertise in reading, with facilitation by CTB/McMillan/McGraw Hill. Task writers were familiarized with the construct underlying the reading test, similar to that of the National Assessment of Educational Progress in Reading (National Assessment Governing Board, 1992), which regards reading as a dynamic process of constructing, extending, and examining meaning. Tasks were developed to provide measures of reading for literary experience, reading to be informed, and reading to perform a task. While attending to these outcomes, those developing MSPAP reading assessment activities used a taxonomy based on research conducted by Langer (1989, 1990). The activities endeavored to stimulate students to engage in four stances, or orientations in response to each reading text: global understanding, developing interpretation, personal response, and critical stance. A wide range of passage elements provided the focus for reading activities; these included, but were not limited to, theme, plot problem or conflict, resolution, major events or main topic, subordinate topics or ideas, purpose, organization, setting, character, style, literary devices, authorial intention, illustrations and adjunct aids or graphics.

In 1991, most reading scoring tools were developed only after activities had been approved by members of an assessment development coordinating team and were no longer subject to

revision. The exception was a set of stance-based reading rubrics which were to be used with extended, prompted written responses to reading passages; these rubrics were drafted early on, and underwent considerable revision after activities were finalized, each retaining, however, an emphasis on one stance or orientation to the text. Working from information on intended stance and with task writer-generated "exemplary responses" under MSDE guidelines, the scoring contractor developed scoring tools for all other reading activities. Some of these tools were considered "rules"; these were scales with brief descriptors that could be generalized across a number of activities (in effect, "mini-rubrics"). The remainder of the tools were activity specific keys--scales with descriptors that were unique to the particular reading activity to be scored. Each tool was accompanied in the scoring guide by answer cues which indicated the possible text information in a complete and acceptable response. Care was taken to emphasize that these cues were not "answer keys"--that is, that any text-based response that addressed the question was scorable, whether or not it was modeled in the guide. Many scoring tools were accompanied in the guide by anchors to illustrate possible responses at each score point. Scoring training also included an introduction to guidelines to assist readers in making decisions, particularly when faced with idiosyncratic or unexpected responses.

Such responses must be expected in any sort of open-ended assessment situation. Since negotiation or assisted

interpretation is not possible in testing contexts, the opportunities for selective readings, unexpected interpretations and misinterpretation are many, and the consequences, at least in terms of inaccurate or inappropriate score decisions, are great. In a performance assessment of reading, the problems of miscueing go beyond the test task to text context, just as they do in writing assessment (Goldberg, 1990); these problems are further compounded, however, by the navigation required by the test taker between the text of the reading passage, the text of the task, and the text that is generated as a response.

There was recognition, even prior to administration, of the likelihood of some miscueing and crosscueing. However, the dynamics of task writing workshops involved people with a range of expertise and notions about assessment, all exploring new assessment territory. As a result, some questionably worded activities to make their way into the final version of student response booklets. Initial perceptions of task features likely to trigger problematic responses were confirmed during the development of scoring keys and answer cues to accompany each brief, open-ended reading item, together with the experience of reading sample responses during rangefinding. These perceptions provided the impetus to go forward with our investigation following the test administration. Through examination of response samples and inquiry about scorers' experiences of unexpected or idiosyncratic responses, we aspired to learn more

about mismatches between the range of intended responses and those actually generated.

Method:

To conduct this investigation, we focused on the two major test forms in reading/writing/language usage out of six forms administered at each grade level (3, 5 and 8). For response samples, we drew upon those in the bank of eighty test booklets that were randomly selected and prescored by reading and writing specialists for purposes of providing training and qualifying sets, as well as conducting validity checks on readers. Each test form contained approximately twenty-six brief open-ended questions (based on three to four authentic texts) and one extended (prompted) question meant to yield a more developed response to be scored for reading, writing, and language usage. Approximately six of the brief reading activities contained a standard cue signalling students that responses to those activities would also be scored for language usage.

In conducting our investigation, we also utilized questionnaire data on these two test forms collected towards the end of the scoring session from all twelve teams of trained Maryland teachers serving as readers (see Appendix A). The data from reading teams represented the responses of 123 teachers. The respondents identified and discussed items the wording of which appeared to limit or adversely affect student responses

and/or scoring decisions. These data were examined in light of 1) evidence of miscueing and crosscueing that had already surfaced from item review prior to and during scoring training and 2) close reading of the sample student responses.

In our initial approach to the "misfires" among reading responses, we found it helpful to apply to MSPAP reading activities the system of classification of response error developed for the field of survey research by Cannell and Robinson (1971) and adopted by Ruth and Murphy in their examination of writing tasks (1988). Thus, reading activities and their associated student responses were examined for evidence of instrument errors (those that are inherent in the task itself), respondent errors (those that occur due to student misreadings of the task) and contextual errors (those that occur because of the discrepancies between oral/informal response to reading texts and written/test-engendered response). While simply a starting point, this schema allowed us to think more systematically about problematic responses and their sources.

Careful reading of the sample responses and application of the corresponding scoring tools, answer cues and anchors led to the identification of those activities which were the occasion of response "misfires." No strict quantification of instances of response error was used to ascertain the point at which an activity should be considered "problematic"; Usually, more than two or three instances of "misfiring" caused us to flag the

activity, and then to check these impressions against the input from the teacher-scorers who had reacted in the questionnaire to activity features near the end of the scoring session. We began to see patterns in cueing underlying various activity-types with problematic responses. Based on resemblances among activities, we found we could predict those singled out by the trained scorers as sources of misfires based on their more extended experience, even if the training pool yielded no problematic responses to those activities. We found that the error sources were finite, falling into general "families;" furthermore, that virtually every one was identified (and thus validated) by the readers through their questionnaire responses. Overall, we were able to identify well over a dozen response error sources, either instrument-based, respondent-based, context-based, or some combination of these. The "softness" of our classifications reflects the fundamental purpose of this activity--not to quantify but to identify problems in the wording or format or reading performance tasks in an effort to reduce or even eliminate such problems in the future.

Findings:

Across the test forms examined, contextual errors--those triggered by the discrepancies between classroom and testing contexts-- were rare. In a few activities, however, there was evidence of cueing that triggered reliance by some students on an implied, shared universe of text information which sometimes led

to incomplete responses, and thus, to lower scores; we considered this a distinct error source. This problem was apparent especially in questions that cued for personal response, such as the following two:

Think about how Eric's friends helped him in the story, and then explain below how an experience of yours was like or unlike Eric's. (Note: because of test security, actual questions have been slightly modified to mask the related reading text but still mirror the format and wording of the particular question).

When you were young, someone took care of you the way Mary cares for Joey. Tell one way you and Joey were both cared for in the same way.

Some students simply identified analogous or contrasting experiences (i.e., "I was sad when I fell" in response to the first activity or "my Mom cuddled me" in response to the second) but provided no evidence of constructing or extending meaning. Their responses were feasible but not text-based ("the same bad thing didn't happen to me as him"). Activities of this sort also triggered responses based on implicit connections to text, assumed from a context in which the student's classroom community had read and shared the same text. We identified this as a context issue since a more "scorable" response could be drawn out in classroom discussion in a way not permitted on a test. We recognize, in retrospect, that cueing for more explicit

comparison/contrast between personal ideas or experiences and text-based information might have been advisable.

This kind of response error also occurred with multi-part questions such as the following:

The process of obtaining things to eat in the country is different from obtaining things to eat in the city. Which do you think is best? Explain your answer on the lines below, using ideas from the story.

Dealing with these sorts of responses created considerable difficulty for scorers who used a rule which granted partial credit to "a relevant personal response" and full credit to a response containing text-based plus personal ideas or experiences. (See Appendix B) Readers struggled with giving partial credit to a personal opinion the explanation for which might be missing or ungrounded in the text. From a scoring perspective, this sort of problem could be addressed by distinguishing between opinion or choice as evidence of personal response vs. personal ideas/information, and only giving partial credit to the latter. From a task development perspective, combining directives in the activity would have addressed this problem, for instance by asking "based on the story and your own ideas, explain which process of obtaining food you prefer, and why."

While its source was not always clearly a contextual one, the problem of feasible, though not text-based, response was not uncommon among responses to 1991 MSPAP reading activities, as it has been in other reading assessments such as the IEA International Reading Literacy Study (see Kapinus, Atash, Binkley and Coley, 1991). At least two or three times in each test form, questions appeared which occasionally generated responses that were legitimate, but gave no evidence of constructing or extending meaning. This error type was relatively more frequent in response to activities intended to provide a measure of reading to perform a task--in instances when the function of a particular text such as a recipe or schedule is evident without reading. For examples, several test forms contained questions such as, *"What was the main purpose of this recipe?"* Scorers were stymied by responses like "the main purpose of the recipe is to teach someone how to cook something." The same problem was also apparent in some critical response questions intended to provide a measure of each of the various purposes for reading; students could rely upon what they know about how texts in general work to answer a question like "would story be as effective without illustration?" by indicating "no, the illustrations help you visualize what's happening." Equally problematic responses were often given to questions about the effectiveness of the title or some other text feature.

In some cases, cueing intended to trigger a particular stance (with scoring to be based on evidence of that stance)

backfired and triggered a feasible but non-scorable response. For example, in response to the question, *"How did the illustrations help you to understand the process?"* some students engaged in a personal stance and responded, "they didn't, since I already knew how to build a doghouse." In contrast, some questions intended to trigger global understanding and to be scored for evidence of synthesis of ideas from the text (*"Write a sentence to tell the author's purpose in writing this article"*) triggered responses dealing with the author's craft. Efforts to meet test specifications in terms of scope of form of activities also backfired on occasion. For instance, some activities intended to prompt the linking of two passages failed to adequately cue students to do so. So, for example, *"Would the protagonist in story A have approved of the protagonist in story B?"* was evaluated for evidence from both texts to explain the answer based on the task-writers' intent, causing an over-abundance of partial credit decisions.

Other kinds of problematic responses were much more clearly caused by the instrument alone--the language and word order of the activity--rather than by some interaction between the activity and testing context. Among the most common instrument error sources was the conflict between reading and writing task cues. In a number of activities, for example, students were given instructions to write a specific number of sentences, to write a paragraph, or a note. Syntactic or rhetorical demands appeared to override cues that would lead to evidence of

interaction with the reading text or texts. A similar conflict could be seen when students were given the opportunity to respond either in a drawing or in words. Literalists that they sometimes are, students (particularly the younger ones) limited themselves to responding "in words," not in more meaning-laden phrases or sentences, so that truly, one picture was worth a great deal more than the verbal conveyance of idea or impression.

Some activities were made problematic by a lack of clarity about the quantity, extensiveness, or uniqueness of response information sought. This occurred for instance when students were asked to "complete the chart," one which provided space but no markers for number of ideas or references. Sometimes students would synthesize response information, so that scorers seeking a certain number of ideas of pieces of information were unsure how to treat those responses. Thus, the student who responded to the request to *"Fill in the graphic organizer to explain how the author shows that John is disobedient"* with one entry--"he never listens to his mother when she asks him to do chores around the house"-- might not get as much credit as if he used two entries--"he doesn't take out the garbage" and "he doesn't do the dishes."

The occasional choice of verbal or graphic response was another instrument feature which sometimes caused problems. When activity directives allowed for a drawing instead of a written response, a "main idea" cue could lead to a very generic drawing that was based on the title alone, or on illustrations already

provided with the passage. Sometimes the concept being sought in a test question was very hard to represent graphically (i.e., what is seen vs. what was felt or understood). So, for example, *"What happened to Mr. Smith at end of story"* is difficult to represent if Mr. Smith has just realized that he has been mortally wounded. Neither the figure of a man standing or lying on the floor accurately conveys what happened.

Inter-item cueing, a pitfall familiar to developers of multiple choice items, led to other response misfires and scoring problems, albeit infrequently. Some test activities contained internal cues for the intended response; others cued for answers to other activities in the task. In some instances, inter-activity directions or bridging material cued students for responses. Scorers were repeatedly frustrated in these situations, since answer cues and the scoring tools themselves were based on intended information, and did not account for information invalidated by inappropriate cueing in test materials of various sorts.

Many of the response misfires we saw were triggered by ambiguity in the wording of activities, another error source that can be traced to the measurement instrument itself. Ambiguity was apparent, for example, in one question attempting to ascertain students' understanding of change. Students were asked to explain how the character *"changed between the beginning and the end of the story"* instead of how he *"changed from the*

beginning to the end of the story." Since the intent was to capture a process of transformation and not a specific change, scorers experienced considerable dissonance when responses did what the activity asked them to do and not what the activity was intended to have them do. Other occasions when activities revealed a lack of clarity included:

1) Cueing indescriminately for examples, text support, text reference, or text relevance. These cues were sometimes used interchangably, while score decisions were based on only one or another of these types of support.

2) Referring to text features like an insert without making clear which components (i.e., text, illustration, and/or the caption) was meant. This created issues in scoring, since information permissible in a student's response depended on an understanding of what the insert included or how it was to be defined.

3) Unintentionally using words with multiple meanings like "common" (typical vs. shared) or "point" (moment in time vs. idea).

4) Asking for an explanation of a double meaning when what was desired was an explanation of both meanings.

Cross-cueing, another instrument error source, was manifest in a wide number of ways. Sometimes, for example, the activity format was such that a question was first posed in one paragraph, and then a lead-in was provided that cued for information different from that cued for in the question itself, as in the instance below:

Look at the list of sports below. Put an X in front of the sport that you feel presents a problem similar to Mark's. On the lines below, tell why you chose that sport and how it is similar to Mark's sport.

/ / skateboarding
/ / bicycling
/ / roller-skating
/ / football
/ / other _____(Name the sport)

How this sport is similar to Mark's: (Lines for response follow)

Cross-cueing also occured when activities asked for both singular and plural items of information ("*tell what character did. He did these things*"). Depending on which words or phrases the student focuses, his response might only receive partial credit because incomplete, having conveyed only one thing the character did.

Students' problems with activity directives frequently highlighted cross-cueing which had them going in directions quite different from those intended. In one of the most extreme instances of this problem, one activity cued students for information within and outside of text, and then waffled over whether one, the other, or both were being sought. Students were first asked to *"Consider the way the poem contrasts (A) and (B). What are the differences between (A) and (B)? Fill in the lines below to show how the poem contrasts (A) and (B)."* Then, in the graphic beneath, students were cued for *"two ideas in the poem that contrast (A) and (B)."*

Problematic responses to MSPAP activities ought not by any means always be attributed to flaws in the instrument, for we certainly encountered instances of true respondant error. Some unexpected responses to test questions were clearly the result of the respondent not understanding, or overriding, the cues provided. In several instances, for example, readers overrode cues about point of view, disregarding instructions to *"imagine you are X. Tell what happened from your point of view."* Some readers in those instances focused on *"tell what happened"* and only received partial credit then, for global understanding without any extension beyond the information in the text.

In another sort of respondent error, sequence cues overrode object cues. When students were asked to tell what happened to X at the beginning, middle, and end, they often misread and skipped

the focus on X. The most obvious of respondent error type, however, might be the occasional incomplete or incorrect reading of a test question. Some students overlooked the phrase with which a question ended (i.e., before X happens), picking up only action and quantity directives that preceded it. Similarly, students failed to attend to a culminating instruction to *"give examples from text"*--some students omitted examples, but wrote wonderful answers to the question itself, many of which were not scorable because there was no evidence of interaction with the reading text, a situation which was far more frequent in a context discussed previously.

Students proved particularly prone to misdirection when activities were intended to stimulate readers to engage in a critical stance towards the text. They sometimes misread questions like *"how does insert provide evidence"* as *"what evidence does the insert provide,"* and only got partial credit. True misunderstandings of the activity, however, were surprisingly, and gratifyingly, few. One of the most amusing and unexpected of this type of error came in response to a question intended to focus on the epistolary nature of one passage. When asked what effect on readers presenting *"letters to pops"* might have had, a number of students responded that writing to pops was not as clear as writing to grandpa or grandfather.

Sometimes we felt it hard to distinguish between true respondent error and response error triggered at least in part by

the format or wording of the activity. This was the case, for example, when there was confusion between what the author did and how he did it. While sometimes this was a clear case of misreading, other times the students were lead to mimic author (do as she does) instead of talking about what author does. Other response errors occurred when there was bridging to new text. In attempting to build a rationale for the introduction of new text, distractors were sometimes inadvertently thrown in. In consequence, readers identified unintended connections between texts. Students shifted their attention to test text, not reading text.

Another error that seemed to have its source not only in the instrument, but also in the respondent's reading of that instrument, involved word elaboration items (those which asked *"which word best describes...explain"*). These sorts of questions frequently caused solipsistic responses. Particularly when a list of adjectives was given to describe a character, followed by the request to explain their choice, students responded by saying X is (adjective) because he (verb form) (i.e., "Jodie is playful because he plays a lot"). These activities were also problematic when a reason for word choice was asked for, rather than example supporting that word choice--nothing cued for expansion or multiple ideas.

Not unexpectedly, word order sometimes caused misfires, another response behavior that appeared to implicate both

instrument and respondent. When a series of directives were given, students tended to read the first directive only or to read the initial portion of questions and jump into the response before the question was fully read or comprehended. Some students also had difficulty because of "word overload." Particularly in activities scored for language usage as well as reading, in which a realistic context for writing was required, the activity became so detailed that some students lost track of the essential question, as in the case of the following, when the students summarized the story but neglected to focus on Christina:

Imagine that you have been asked to share the important parts of the story in a book report for the bulletin board. As part of your book report, write what you remember about Christina on the lines below.
Because your friends will read the report, you will want to do your best work. Be sure to check for correct spelling, grammar, punctuation, and capitalization. Make any necessary corrections neatly.

Other times, when an activity began with "*think about X, think about Y*" and then finally got to the essential question, students lost focus. Complex cognitive equations seemed to be set up that got in the way.

Practical Outcomes

Based on these findings from the 1991 MSPAP, our increased awareness of the problems causing mismatch between activity and intended range of response has significantly affected efforts aimed towards the development and scoring of the 1992 MSP. Having recognized and identified models of various error sources, we were able to implement some immediate changes and additions to the task development process. One, in particular, was the introduction of a scoring review checklist, for use during test development, which identified potentially problematic features of open-ended activities and gave examples of undesirable response behaviors attributable to each feature. During preliminary drafting and review, task writers were sensitized to potential pitfalls and were able to eliminate many quite early on in the development process. This checklist was based on problems apparent not only in reading activities, but in each of the content areas assessed in 1991. In turn, it was found useful not only in the development of reading, writing/language usage and mathematics activities, but in science and social studies, the two new content areas added to MSPAP in 1992. The introduction to scoring-related issues by means of applying the checklist was followed by the training and folding-in of reading "scoring team" members drawn from teacher-scorers from the 1991 project to the 1992 task development teams.

Concurrently, the reading rubrics and scoring methodology were reexamined to reduce the artificial constraints in terms of

reading stances that sometimes existed in 1991. The stage for this was set with the reevaluation by task writers of the role of stances in new activities being created for the 1992 test. Instead of forcing a one-to-one correspondence between activity and intended stance, task developers tried to better anticipate response possibilities in terms of more complex movements within and among stances. A generic rubric was devised to capture the movement through multiple stances likely in a successful response to a more complex question. (see Appendix C) Trials were conducted in designing scoring rules and keys that relate more closely to the generic rubric, and thus, to the reading outcomes being measured, instead of reflecting a workbook score-key strategy. In both activities and associated scoring tools, the emphasis is now much more clearly on engendering and evaluating construction, extension and examination of meaning. It is hoped that not only will the new scoring tools be more accurate and efficient when used for scoring but that they will also impact the analysis and reporting of results and the application of the performance criteria by students and teachers.

At the same time, other efforts have been made to impact instruction--not by stimulating teaching to the test, but to make clear the importance of looking to the text for evidence to support ideas and opinions. It is hoped that over time, it will become less necessary to specifically cue for text support, once students recognize this as an essential part of responding to reading questions.

Nevertheless, meanwhile, scorers will be provided with clearer guidelines on the need for evidence from the text in making score decisions. Decision-making in scoring should also be facilitated by the introduction of a newly defined condition code. Non-scorable, long the designation for responses that are off-topic, off-task, will be used instead to identify responses that cannot be scored because while they are feasible, they provide no evidence of the outcome being measured. Tracking the frequency with which this code is assigned to particular items will help to cull ineffective activities from the assessment prior to analysis and reporting of test results.

A central aim in any direct assessment is to ensure valid and reliable test items/activities. Especially given the recent and increasing emphasis on performance assessment in reading, it is likely that even with conscientious review during task development, some problematic items will surface upon test administration. The process of task, scoring tool, and response analysis coupled with trained reader review provides a means of identifying characteristics of performance tasks which may lead, directly or indirectly, to problematic responses. It is hoped that this study will contribute to the refinement of task development and scoring by other programs considering or already engaged in performance assessment. Further, it should be noted that some respondent error can be eliminated through effective instruction that develops students' ability to carefully read and

respond to performance assessment tasks. As assessments become more performance-oriented and tasks become more complex, students' proficiency in these areas will become an essential outcome of education, making this study meaningful from instructional as well as measurement perspectives.

References

Cannell, C. F., and Robinson, S. (1971). Analysis of individual questions. In J. B. Lansing, S. B. Withey, and A.C. Wolfe (Eds.), Working papers in survey research in poverty areas. Ann Arbor, MI: University of Michigan, Institute for Social Research.

Goldberg, G. (1990). Topic literacy and off-topic responses. Paper presented at the Eighth Annual Conference on Language and Literacy Assessment (National Testing Network in Writing), New York, NY.

Kapinus, B., Atash, N. and Binkley, M. (1991). Constructed response items in large-scale reading assessment: plus, minus, and interesting. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago.

Langer, J. A. (1989). The process of understanding literature. Report Series 2.1. Albany, NY: Center for the Learning and Teaching of Literature.

Langer, J. A. (1990). The process of understanding: reading for literary and informative purposes. Research in the Teaching of English, 24, 229-260.

National Assessment Governing Board (1992). Reading Framework for the 1992 National Assessment of Educational Progress. Washington, D.C.: National Assessment Governing Board.

Ruth, L. and Murphy, S. (1988). Designing Writing Tasks for the Assessment of Writing. Norwood, N. J.: Ablex.

G19:aerapr3

test form _____ module number _____

today's date _____ # of team members responding (including module leader)_____

READERS' QUESTIONNAIRE

(Reactions to test characteristics that affected scoring)

The following questions are to be discussed among scoring team members and a consensus response is to be recorded by the module leader of that team:

1. In the module you scored, were there any items the wording of which appeared to limit or adversely affect student responses? Give item number(s) and explain:

2. In the module you scored, were there any items the wording of which limited or adversely affected your score decisions? Give item number(s) and explain:

3. In the module you scored, were there any scoring tools (keys, rules, rubrics) the wording of which limited or adversely affected your score decisions? Give item number(s) and explain:

(Please complete second page)

4. In the module you scored, were there any items the format/graphics/layout of which appeared to limit or adversely affect student responses? This may apply to either the question itself or the area provided for the student response. Give item number(s) and explain:

5. In the module you scored, were there any items the format/graphics/layout of which limited or adversely affected your score decisions? Give item number(s) and explain:

6. In the module you scored, were there any items the format/graphics/layout of which adversely affected the ease of scoring? Give item number(s) and explain:

(Please use additional page(s) if necessary)

Appendix B

Outcome: Information
Stance: Personal Response

Answer Cue	Points
Personal experience linked to 1 of the following: ▪ both were loved/given affection ▪ both were fed ▪ both were kept clean ▪ both were held/cuddled	2 = relevant personal experience plus one text-based idea 1 = relevant personal experience or one relevant text-based idea, but not both 0 = other

Outcome: Information
Stance: Developing
Interpretation

Reading Scoring Rubric: Developing Interpretation

3 Points: These responses demonstrate evidence of an extensively developed interpretation of the text. These responses contain both inferences grounded in the text and supporting information from the text. The inferences clearly extend beyond the information in the text and are linked to the text by explicit, substantial references.

2 Points: These responses demonstrate evidence of an adequately developed interpretation of the text. These responses contain some inferences grounded in the text and some supporting information from the text. Some inferences extend beyond the information in the text and are linked to the text by some references.

1 Point: These responses demonstrate evidence of a partially developed interpretation of the text. These responses either do not move beyond the literal information in the text or offer inferences without text support. These responses may contain both correct and incorrect information from the text.

0 Points: These responses demonstrate evidence of an attempt to respond to the question. These responses do not contain evidence of understanding the text. They contain either incorrect or irrelevant information. These responses may contain only sections copied from the reading text.

Outcome: Literary Experience
Stance: Personal Response

Reading Scoring Rubric: Personal Response

3 Points: These responses demonstrate evidence of an **extensive** personal response. These responses contain both personal ideas related to the text and a clear presentation of the meaning of the relevant parts of the text. Where called for, these responses are substantially supported with personal experience/opinion/judgement and/or references to the text. The link between the text and personal experience/opinion/judgement is relevant and explicit.

2 Points: These responses demonstrate evidence of an **adequate** personal response. These responses contain both personal ideas and some representation of the meaning of the text. Where called for, these responses are somewhat supported with personal experience/opinion/judgement and/or references to the text. The link between the text and personal experience/opinion/judgement is relevant but may not be explicit.

1 Point: These responses demonstrate evidence of a **partial** personal response. They contain either personal experience/opinion/judgement or reference to the text but not both. If both types of information are present, the text information is partially incorrect or irrelevant. The attempted link is either not relevant or not explicit.

0 Points: These responses demonstrate evidence of an attempt to respond to the question. These responses do not contain evidence of either understanding the text or making a personal response linked to the text.

Outcome: Literary Experience
Stance: Critical Stance

Reading Scoring Rubric: Critical Stance

3 Points: These responses demonstrate evidence of an extensiv consideration of the text through a critical stance that examine the author's craft and purpose. These responses contai inferences about the author's purpose or background or identif elements of the writer's craft that the author uses to develo the text.

- If the response requires inferences about the author's background or purpose, the inferences are clearly and substantially supported by references to the text.
- If elements of the writer's craft are identified, their contribution to the text is clearly and substantially discussed.

2 Points: These responses demonstrate evidence of an adequat consideration of the text through a critical stance that examine the author's craft and purpose. These responses contain inferences about the author's purpose or background or identif elements of the writer's craft that the author uses to develo the text.

- Inferences are sound and supported by some reference(s) to the text, or;
- Elements of the writer's craft are identified and their contribution to the text is stated.

1 Point: These responses demonstrate evidence of a partia consideration of the text through a critical stance that examine the author's craft and purpose. These responses may:

- contain inferences but fail to support them, or;
- fail to give information on the author's purpose or craft, although there is an indication of an understanding of the text elements such as humor or suspense, or;
- give both correct and incorrect information.

0 Points: These responses demonstrate evidence of an attempt to respond t the question. These responses do not contain evidence o understanding the text or the author's purpose and ways o crafting the text.

BEST COPY AVAILABLE

Appendix C-4 (1992 rubric)

Generic Rubric for Reading Activities

Note: The scoring criteria are applied as cued for or required by the activity. Not every activity cues for all the possible behaviors described in the generic rubric. Thus the scoring guide for specific activities can stop at different points. For some activities, the scale will shift by one point downwards. It is possible in some instances for this rubric to be applied across more than one activity in a task, in which case the scorer would determine which score point description generally characterizes the responses being considered.

0 = No evidence of construction of meaning.

1 = Some evidence of constructing meaning, building some understanding of the text. Presence of defensible, and possibly some indefensible, information.

2 = A superficial understanding of the text, with evidence of constructing meaning. One or two relevant but unsupported inferences.

3 = A developed understanding of the text with evidence of. connections, extensions, or examinations of the meaning. Connections between the reader's ideas and the text itself are implied. Extensions and examinations are related to the text but explicit references to the text in support of inferences are not present. When more than one stance is possible, the response may remain limited to one stance.

4 = A developed understanding of the text with evidence of connections, extensions, and examinations of meaning. Connections between the reader's ideas and the text itself are explicit. Extensions and examinations are accompanied by explicit references to the text in support of inferences. When possible, the response indicates more than one stance or perspective on the text; however, only one stance is substantially supported by references to the text.

5 = A developed understanding of the text with evidence of connections, extensions, examinations of meaning and defense of interpretations. Connections between the reader's ideas and the text itself are explicit. Extensions and examinations are accompanied by explicit references to the text in support of inferences. When possible, the response indicates more than two stances, all substantially supported by references to the text.

6 = A complex, developed understanding of the text with evidence of connections, extensions, examinations of meaning, and defense of interpretations. Connections tween the reader's ideas and the text itself are explicit. Extensions and examinations are accompanied by explicit references to the text in support of inferences. Responses indicate as many stances as possible, based on the activity, all substantially supported by references to the text. These responses reflect careful thought and thoroughness.